

# Máster en Programación avanzada en Python para Big Data, Hacking y Machine Learning

Programación Python para Machine Learning

# LECCIÓN 02

## Estadística descriptiva, visualización y preparación de datos.

# ÍNDICE

- ✓ Introducción
- ✓ Objetivos
- ✓ Estadística descriptiva
- ✓ Visualización de datos
- ✓ Preprocesamiento de datos
- ✓ Conclusiones

# OBJETIVOS

Al finalizar esta lección serás capaz de:

- 1 Entender los datos para obtener el máximo rendimiento de ellos.
- 2 Utilizar las técnicas de estadística descriptiva para resumir los datos.
- 3 Analizar las relaciones presentes en los datos, numérica y gráficamente.
- 4 Conocer los principios y saber aplicar las técnicas de preprocesamiento de datos.

# INTRODUCCIÓN

- ✓ Conocer los datos:
  - Descubrir las relaciones entre variables.
  - Sesgo.
  - Balanceo de clases.
- ✓ Herramientas:
  - Estadística descriptiva.
  - Visualización de datos.



## UN VISTAZO A LOS DATOS



Nada mejor que un vistazo de los datos en bruto.

Dimensiones del problema.

Tipos de características

Los términos: instancias, patrones, puntos, observaciones, registros, filas... se refieren conceptualmente a lo mismo, cada uno de los datos de los que se disponen para hacer un análisis.

De manera análoga, los términos: características, factores, dimensiones, variables, atributo, propiedad, campo, columnas... son los atributos que describen cada una de las instancias del conjunto de datos.

### Estadísticos descriptivos

- Conteo.
- Media.
- Desviación típica.
- Valores máximo y mínimo.
- Q1, Q2 (mediana) y Q3.



## Desequilibrio de clases

Muchas más observaciones para una clase que para otra.

Ratios:

1:10, 1:100, 1:10<sup>3</sup>, 1:10<sup>4</sup>...

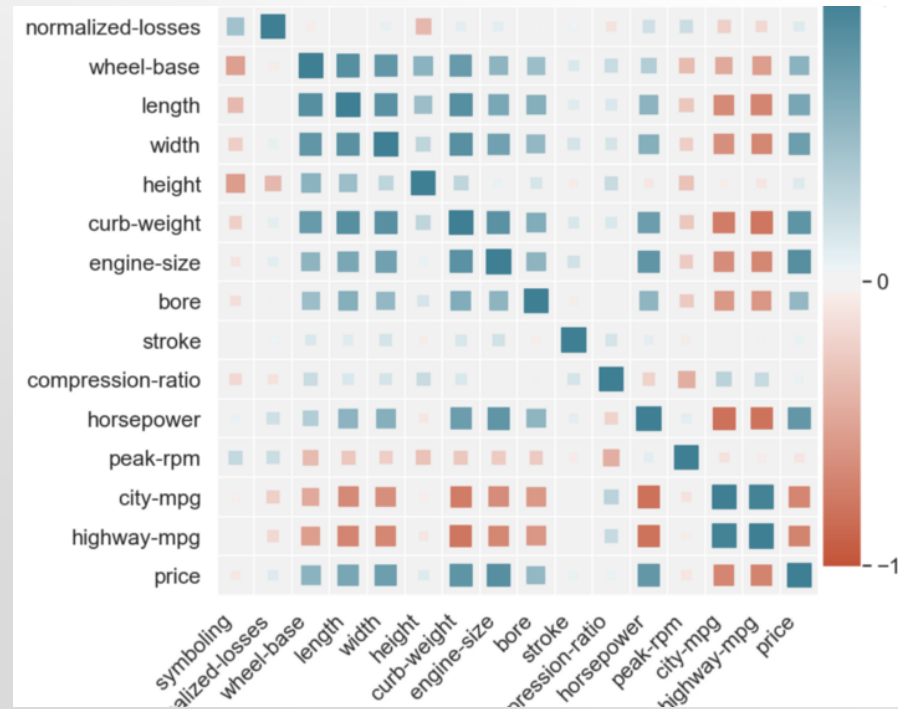




## Correlación de atributos

Relación entre dos variables y de cómo cambian al mismo tiempo.

Problemas de algunos métodos de ML con atributos correlados.



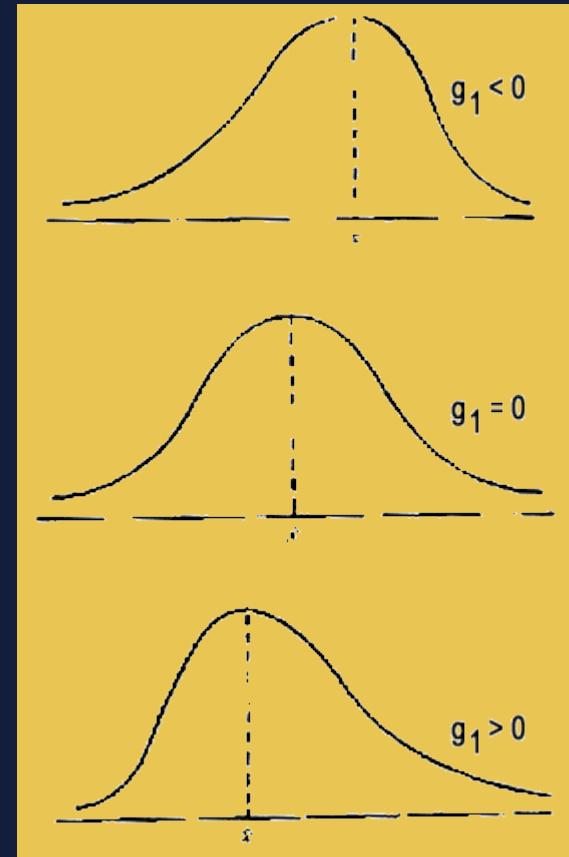
## Sesgo de la distribución

Simetría en la forma de la distribución.

Mayor, menor o igual a 0.

Coefficiente de Fisher:

$$g_1 = \frac{\sum_{i=1}^N (x_i - \bar{x})^3}{N \cdot S_x^3}$$



## VISUALIZACIÓN DE DATOS

Gráficamente es el modo más rápido de adquirir una idea inicial de los datos.

### Gráficos univariantes

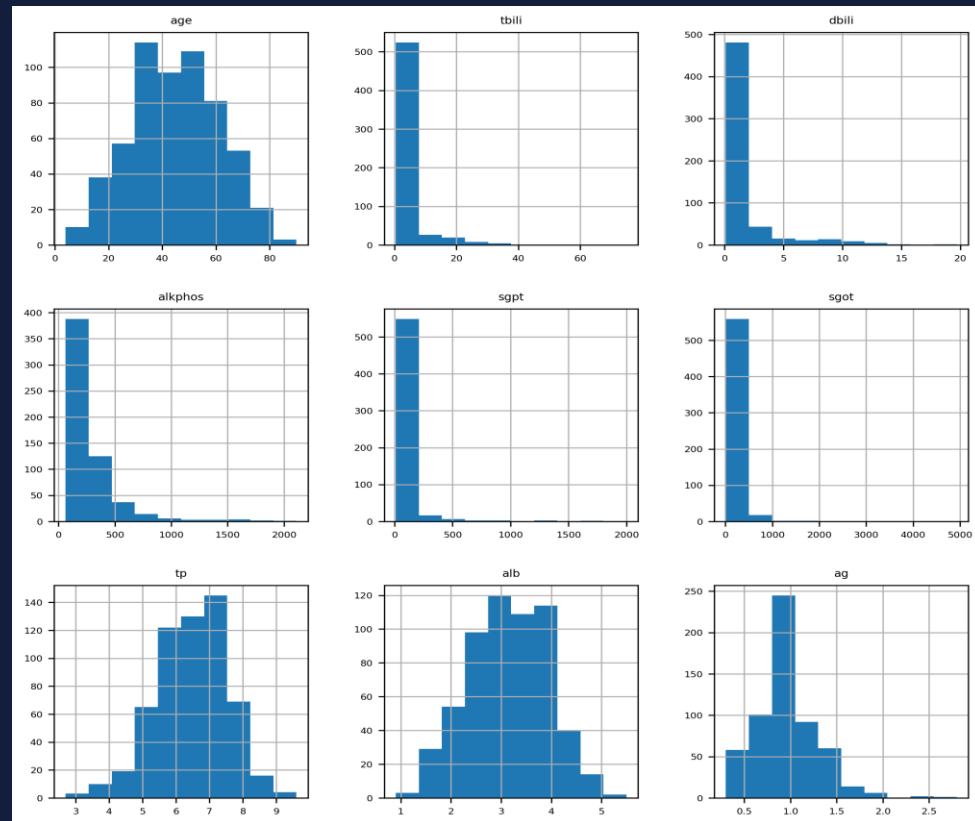
- | Histogramas.
- | Gráficos de densidad.
- | Gráficos de cajas y patas (boxplot).

### Gráficos multivariantes

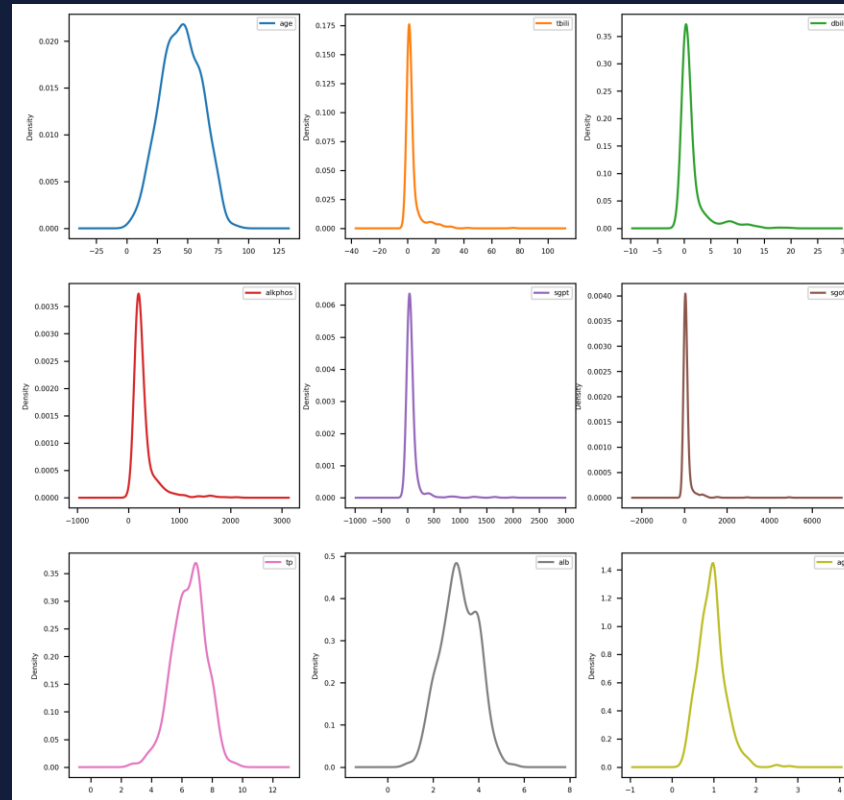
- | Matriz de correlación.
- | Matriz de dispersión



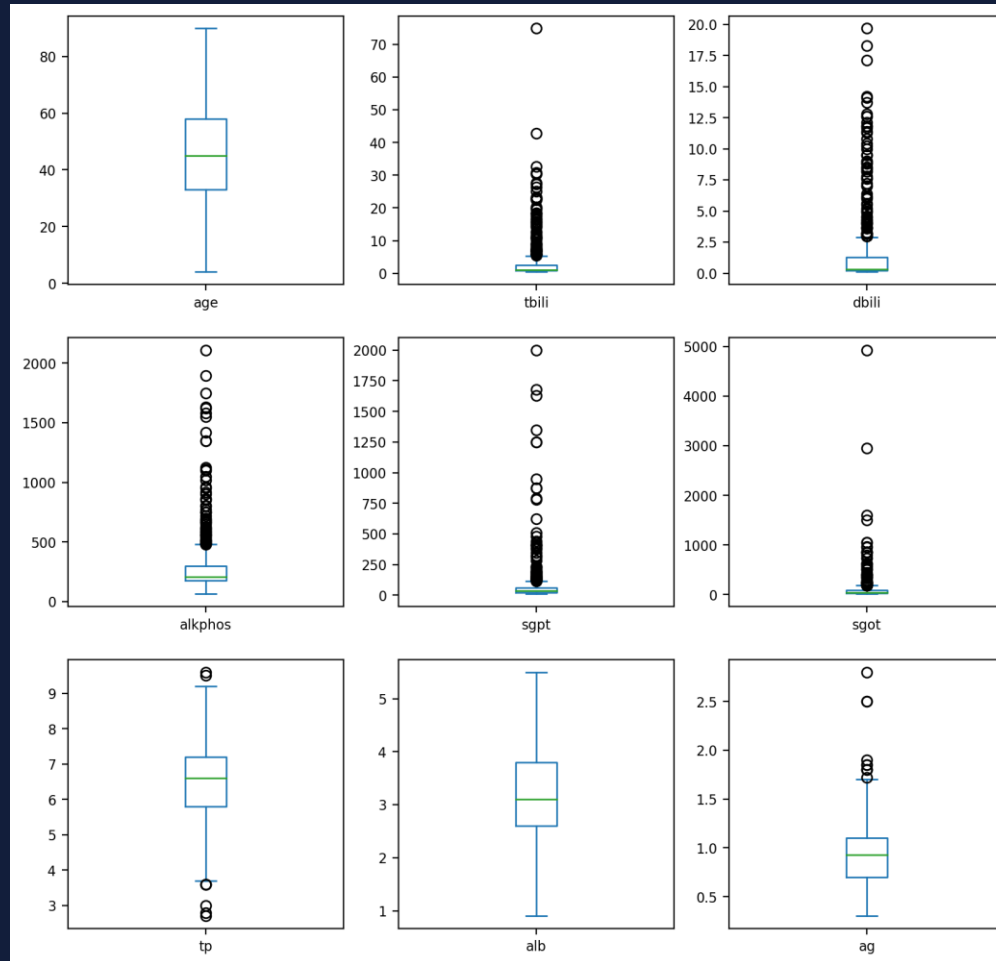
## Histogramas



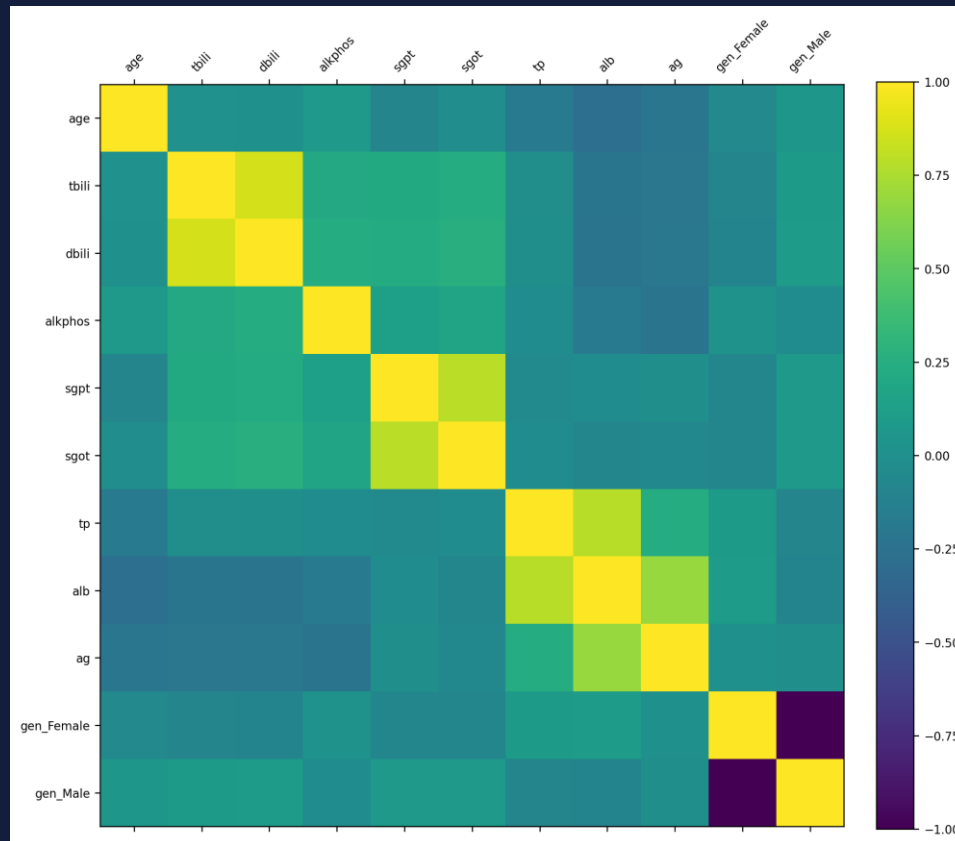
### Gráficos de densidad



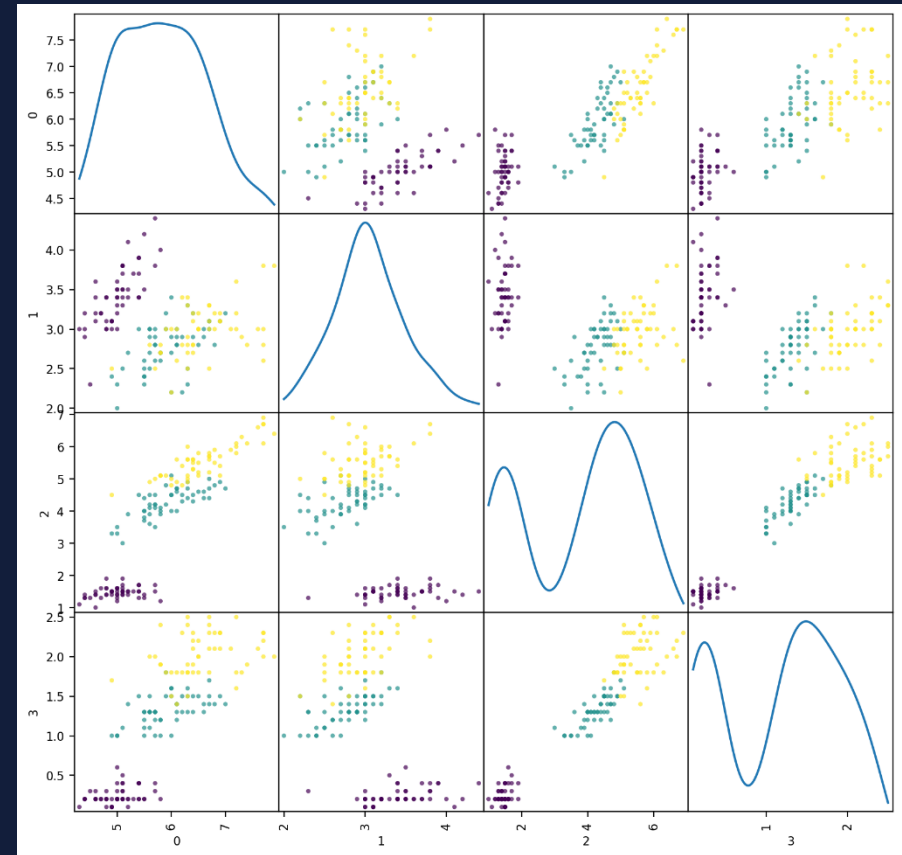
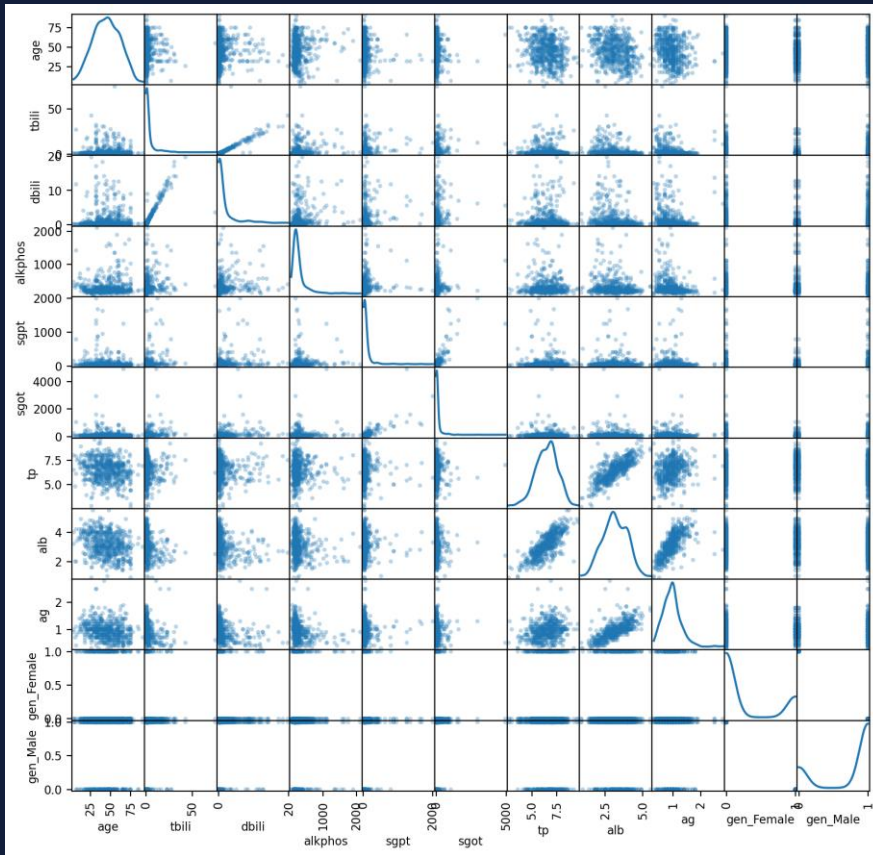
## Boxplots



## Matriz de correlación



## Matriz de dispersión





## TRANSFORMACIONES DE LOS DATOS

Preprocesamiento de los datos para mejor rendimiento de los modelos de Machine Learning

Procedimiento:

- 1) Cargar el conjunto de datos.
- 2) Dividir el conjunto en las variables de entrada y objetivo.
- 3) Aplicar un preprocesamiento mediante una transformación de las variables de entrada.
- 4) Mostrar el cambio producido.




Sobre el conjunto de entrenamiento

## Tratamiento de datos categóricos

Problemática de los algoritmos de ML a la hora de trabajar con datos categóricos.

One-hot-encoding.

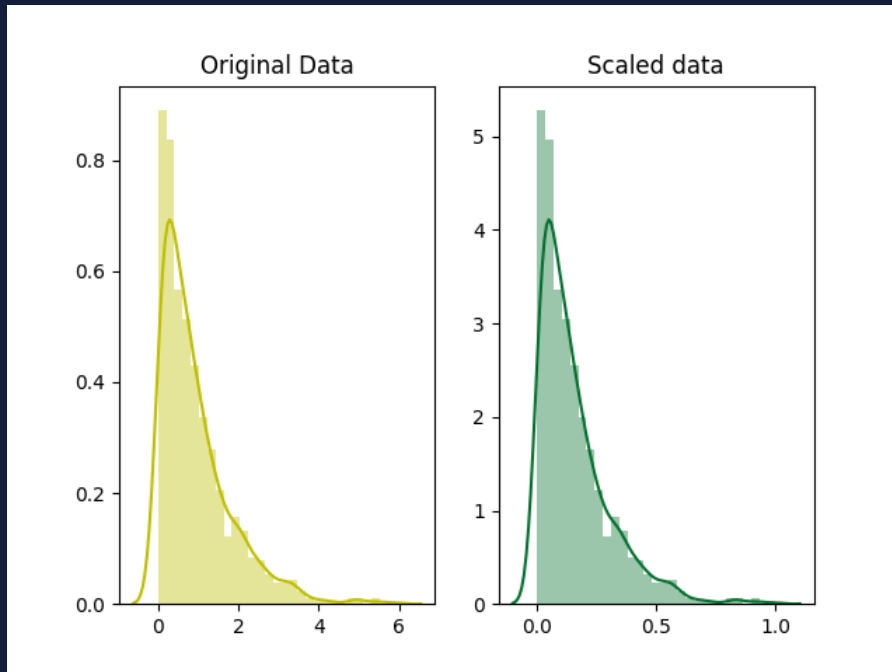
Hay otras alternativas.



Color			
Red			
Red	1	0	0
Yellow	1	0	0
Green	0	1	0
Yellow	0	0	1

## Cambiar la escala de los datos

Problemática de los algoritmos de ML a la hora de trabajar con datos en diferentes rangos.

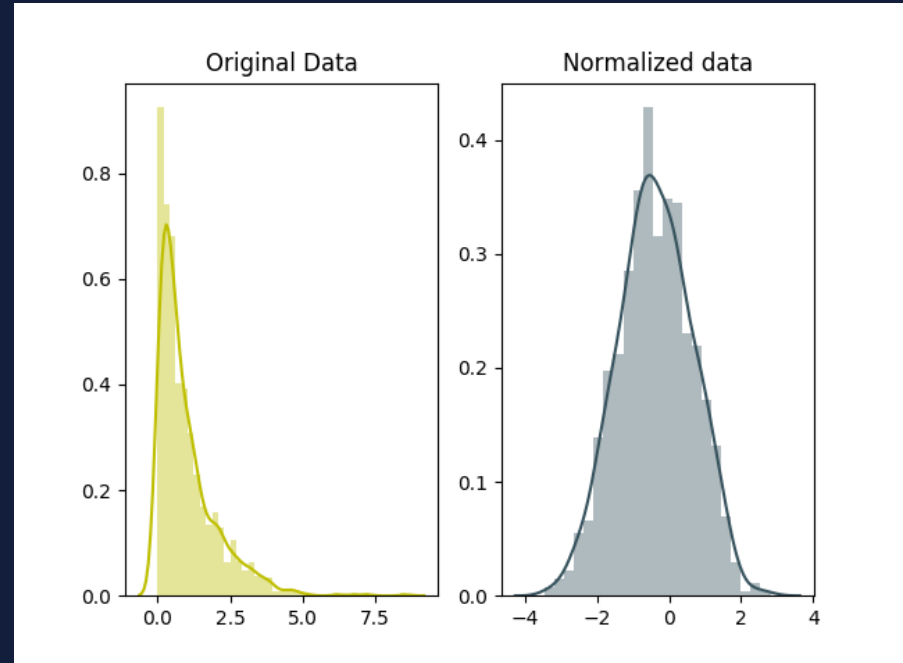


## Estandarizar los datos

Problemática de los algoritmos de ML a la hora de trabajar con datos con distribución no gaussiana.

Estandarización estándar: A una  $N(0,1)$

Estandarización robusta: Datos atípicos.



MUCHAS GRACIAS POR SU ATENCIÓN



[jperez@grupomainjobs.com](mailto:jperez@grupomainjobs.com)



Javier Pérez Rodríguez  
[www.linkedin.com/in/perezxavi](http://www.linkedin.com/in/perezxavi)



[twitter.com/eiposgrados](https://twitter.com/eiposgrados)



[facebook.com/eiposgrados](https://facebook.com/eiposgrados)



[instagram.com/eiposgrados](https://instagram.com/eiposgrados)